(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
30 November 2000 (30.11.2000)

PCT

(10) International Publication Number
WO 00/72013 A1

(51) International Patent Classification[7]: G01N 33/53, 33/543, C12N 15/00, C07H 21/02, 21/04

(21) International Application Number: PCT/US00/13813

(22) International Filing Date: 19 May 2000 (19.05.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/135,429 21 May 1999 (21.05.1999) US
60/172,525 17 December 1999 (17.12.1999) US

(71) Applicant: THE PENN STATE RESEARCH FOUNDATION [US/US]; 304 Old Main, University Park, PA 16802 (US).

(72) Inventors: BENKOVIC, Stephen, J.; 771 Teaberry Lane, State College, PA 16801 (US). OSTERMEIER, Marc; 828 S. Allen Street, State College, PA 16801 (US). NIXON, Andrew, E.; Dyax Corporation, One Kendell Square, Building 600, Cambridge, MA 02139 (US). LUTZ, Stefan, A.; 500 Toftrees Ave. #228, State College, PA 16803 (US).

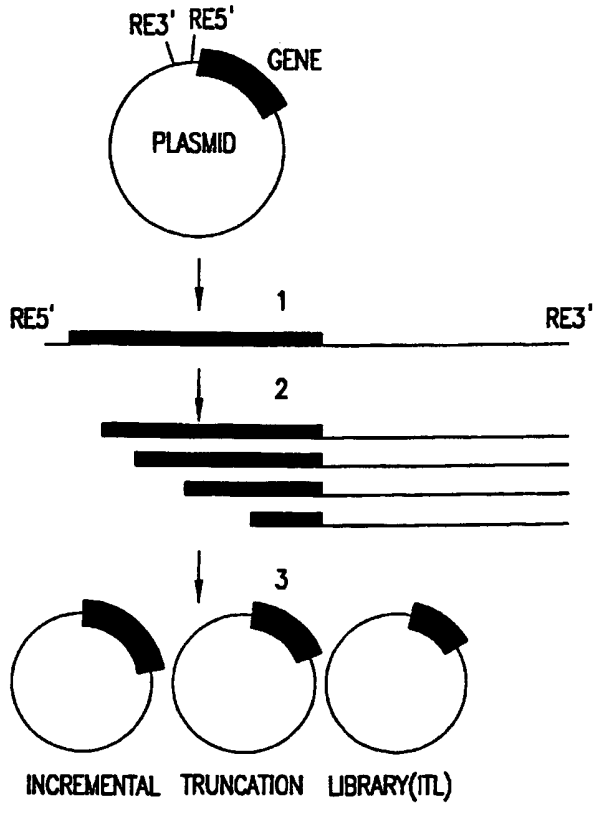
(74) Agent: MONAHAN, Thomas, J.; Intellectual Property Office, The Pennsylvania State University, 113 Technology Center, University Park, PA 16802-7000 (US).

(81) Designated States *(national)*: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian

[Continued on next page]

(54) Title: CONSTRUCTION OF INCREMENTAL TRUNCATION LIBRARIES



(57) Abstract: The application and success of combinatorial approaches to protein engineering problems has increased dramatically. However, curren directed evolution strategies lack combinatorial methodology for creating libraries of hybrid proteins which lack high homology or for creating libraries of highly homologous genes with fusions at regions of non-identity. We have developed a series of combinatorial approaches that utilize the incremental truncation of genes, gene fragments or gene libraries to create such hybrid protein libraries. A library of all possible single base-pair deletions of a given piece of DNA is created. Incremental truncation libraries (ITLs) (as depicted in the figure) have applications in protein engineering as protein folding, protein evolution, and the chemical synthesis of proteins. In addition, a methodology of shuffling ITL's which is independent of DNA sequence homology has been developed.

WO 00/72013 A1

patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

Published:

— *With international search report.*

CONSTRUCTION OF INCREMENTAL TRUNCATION LIBRARIES

This application claims priority of Provisional Application No. 60/135,429 (filed May 21, 1999) and Provisional Application No. 60/172,525 (filed December 17, 1999).

This invention was made with United States Government support in the form of a grant from the National Institute of Health, Grant No. GM24129 and a National Institute of Health postdoctoral fellowship Grant No. GM18560.

FIELD OF THE INVENTION

The present invention is generally directed towards construction of incremental truncation libraries for the creation of hybrid proteins. The present invention provides benefits over previous techniques which required a degree of sequence similarity for the generation of such libraries.

BACKGROUND AND RELATED ART

Protein mutagenesis has long been used as a tool for structure/function studies of proteins. With the advent of modern DNA manipulation techniques and advancements in protein structure determination, large numbers of protein sequences and structures are available which can be sorted into groups or superfamilies based on structural similarity. Such groupings demonstrate that proteins that are structurally similar often catalyze similar reactions and have active sites with shared amino acids. Further, this facilitates identification of side chain residues that are important in binding and catalysis, and allows for their modification so as to yield proteins with altered properties.

Such structure-based rational approaches to protein engineering, through introduction of point mutations, exchange of

secondary structural elements, and exchange of whole domains or subunits, have given rise to enzymes that have altered substrate specificities, catalytic properties and oligomeric states. Although few protein-engineering failures have been published, the difficulty in rationally engineering an enzyme to have a specific function is widely appreciated. Any alteration introduced into a wildtype protein can disrupt the fine balance that nature has achieved, often in unpredictable ways, and consequently give rise to proteins that are unstable, fail to fold properly and lack catalytic activity. As a result of the difficulties encountered using strict rational design approaches, there is an increasing trend towards the use of molecular biology strategies that mimic evolutionary processes. These strategies are known as "directed evolution."

Most directed evolution strategies incorporate some method of introducing random mutations into a gene followed by screening or selection for a desired property. The cycle is then repeated several times until the desired property is achieved or until further cycling produces no improvement in the desired property. Early methodologies utilized point mutations generated by error-prone PCR, chemical mutagenesis or mutator strains of *E. coli*. This type of approach is something akin to an asexual evolutionary process with non-beneficial and beneficial mutations becoming fixed. Such strategies have been particularly successful in achieving improvements in thermostability, altering substrate specificity, and improving activity in organic solvents. However, since directed evolution is a stepwise process, only relatively small steps in sequence space can occur. Thus, the utility of current directed evolution methodologies to evolve novel catalytic sites, which presumably require large excursions in sequence space, is limited.

The advent of methods for recombination, which more closely approximates the natural evolutionary process, has had an enormous impact on directed evolution. In various methods for recombination, such as DNA shuffling, parental genes are fragmented and subsequently reassembled by PCR to reconstitute the full-length

genes. During this reassembly process, novel combinations of the parental genes arise along with new point mutations. This recombination or shuffling approach generates a large library of mutant genes wherein genes that exhibit a desired function can be selected from by using an appropriate selection or screening system.

While it is true that shuffling of families of genes with DNA homology can create hybrid proteins with new properties, such molecular breeding is only feasible for genes with sufficient genetic homology and, for this reason, is unlikely to evolve entirely novel function. It is important to realize that the primary rationale for success in the shuffling of families of genes is the similarity of the three-dimensional structures of the proteins they encode, not the degree of DNA homology. Successful directed evolution on homologous families might be equally or better served by the creation of genes with crossovers between family members at regions of little or no genetic homology. However, current DNA shuffling methodologies only produce crossovers within regions of sufficient homology and within significant stretches of identity. Furthermore, crossovers are biased towards those regions of highest identity.

The increasing numbers of protein structures available and the study of enzyme structural families have shown that many proteins with little or no DNA homology can have high protein structural homology. Constructing hybrids of such structural homologs may well be an important strategy for engineering novel activities; however, no combinatorial approach for the construction of such hybrids has been reported.

The combinatorial approach for the construction of hybrid proteins stemmed from our work in the inter-conversion of formyltetrahydrofolate utilizing enzymes. It has been demonstrated that the feasibility of creating active hybrids between such proteins by engineering a functional hybrid enzyme through fusing domains from two enzymes that overall had very little genetic homology. Discreet domain fusions were made between the glycinamide ribonucleotide

(GAR) binding domain of the *E. coli purN* gene (GAR transformylase) and the formyl-tetrahydrofolate binding and catalytic domain of the *E. coli purU* gene (formyltetrahydrofolate hydrolase). Although a hybrid enzyme was created that had the desired property (GAR transformylase activity), this activity was low prompting a search for a combinatorial approach to this engineering problem as described in Ostermeier, Nixon, Shim, and Benkovic, PNAS USA , 96, 3562-3567 (1999), incorporated herein by reference in its entirety.

SUMMARY OF THE PRESENT INVENTION

Knowing where to make the fusions is a central problem in the creation of novel hybrids. There is a general desire in the field to develop a new approach to the problem of creating hybrid proteins by a method independent of DNA sequence homology between the two genes, gene fragments or libraries of genes or gene fragments being fused. Through incremental truncation methodologies of the present invention, one can create fusion libraries of many (or all) different combinations of lengths of two genes. The novel approaches as described herein are a combinatorial solution to the fundamental questions "where can proteins or protein fragments be fused to produce active hybrids?" and "where are the points at which a protein can be bisected and retain or exhibit desired properties?" Importantly, one aspect of the invention involves various methods that circumvent homology limitations of methods of DNA recombination by allowing genes to be shuffled independent of their sequence homology. Additionally, it should be pointed out that embodiments of the present invention are also useful for creating hybrid proteins from genes with high levels of sequence homology. The present invention, since it is independent of DNA sequence homology, will be applicable to potentially any desired gene, gene fragment, or gene library for the creation of hybrid proteins.

One embodiment of the current invention is directed to a DNA library comprised of a recombinant plasmid incorporating a parent gene and a second recombinant plasmid incorporating a

modified gene. It is preferable that a modified gene in the DNA library have an increment of truncation on the order of about 1 to about 500 nucleotides, but preferably less than 250 nucleotides, even more preferably less than 100 nucleotides, yet even more preferably less than 50 nucleotides, and most preferably the increment of truncation of the modified gene is 1 nucleotide.

In this embodiment, modifications can be made so that the second recombinant plasmid includes a plurality of other recombinant plasmids. These other recombinant plasmids each include truncated versions of the parent gene. The parent gene can be selected from the group consisting of a gene of interest, a portion of a gene of interest, a gene fragment, a PCR product, and/or a mutant of said gene of interest. It is preferable that the modified gene be formed by incremental truncation of the parent gene under conditions suitable to ensure reduction of base pairs at a predetermined rate. It is preferable that this predetermined rate be less than about 50 base pairs per minute and even more preferably less than about 10 base pairs per minute. As in the case where there is a single second recombinant plasmid, when the plurality of recombinant plasmids are used, they can incorporate different modified genes that result from progressive truncation of the parent gene. "Progressive truncation" of the parent gene includes the activity of subsequent removal of nucleotides during the truncation process.

Another method of the present invention is a method of protein identification which includes the steps of isolating a parent gene of interest; truncating that parent gene of interest in a progressive and controlled manner to form a truncated gene; and expressing a protein corresponding to that truncated gene. In the situation where a protein is desired to be identified, the expressed protein is selected based upon the presence or absence of a predetermined characteristic.

It is to be noted that there is a subtle difference between the use of the term "selected" and the use of the term "screening" in the

present invention. The use of selecting looks for a predetermined characteristic which can be, for example, antiobiotic resistance or growth on an auxotrophic host. The use of screening refers to the ability of the resulting expressed protein to exhibit a predetermined activity such as a therapeutic activity or a certain functionality.

As in the previous embodiment of the present invention, it is important in the truncation step that the reduction of nucleotides occur in a progressive and controlled manner, that is, to ensure that relatively small groups of nucleotides are removed during the truncation process. The removal of the number of nucleotides is on a case-by-case basis, but preferably in a range of 1-500 nucleotides, more preferably less than 250 nucleotides, even more preferably less than 100 nucleotides, more preferably less than 50 nucleotides, even more preferably less than 10 nucleotides, and most preferably a nucleotide at a time (*i.e.*, 1 nucleotide). It is the progressive and controlled manner in which the nucleotides are removed from the parent gene that achieves the controlled reduction of the nucleotides as described above. It is best that the conditions be controlled as described herein to ensure that a removal rate of less than 10 base pairs per minute of nucleotides is achieved.

A variety of truncated genes may be used to form a library of proteins which originate from a plurality of differentially modified parent genes. In this case, each member of the library of proteins possess the predetermined characteristic. The parent gene may be originally derived from the gene of interest, a portion of the gene of interest, a gene fragment, a PCR product and/or a mutant of said gene of interest. In this embodiment the libraries are preferably screened or assayed to look for desired activity.

Yet another embodiment of the present invention is a method of producing a protein which comprises isolating a first and a second parent gene; truncating said first and second parent genes in a controlled manner to form a first and a second truncated gene and joining the first and second truncated genes to form a combined

truncated gene. The first and second parent genes by be chosen independent of homology. The term "independent of homology" is meant to connote that the selection process is not dependent on homology between genes. That is, the process will work whether or not a substantial degree of homology exists. However, homologous genes may also be employed and this is not excluded by the phrase independent of homology.

The step of joining may include the step of fusing and/or ligating as described herein. The truncation should be done in a controlled manner which may be either time and/or temperature dependent. Another embodiment includes a method of creating a library of DNA which comprises of initiation a controlled truncation of a parent a gene periodically removing aliquots during the truncation so as to isolate incrementally truncated versions of said parent gene thereby creating a library of truncated DNA. Other embodiments described herein do not require the periodic removal of aliquots during the truncation.

Again, it is important that the rate of removal of the base pairs be less than 100 base pairs per minute, and even more preferably less than 10 base pairs per minute. The DNA construct is usually in the form of a recombinant plasmid and may include any of the original or source recombinant plasmids which may include a functional genetic element which is used to facilitate fusion of the first and second parent genes.

This library preferably includes a plurality of different combined truncated genes which may be used later to express proteins having different characteristics. The library of randomly combined truncated genes is used to express proteins which may have a predetermined characteristic or activity. Therefore, the proteins that are produced by the randomly combined truncated genes can be selected and/or screened to determine the presence or absence of a predetermined characteristic or activity. It is a preferable aspect of the

present embodiment that the selected proteins are screened for activity as well as for the predetermined characteristic.

Although the randomly combined truncated genes or the library of randomly combined truncated genes may be used to produce a library of proteins which then may be randomly combined to produce a library of proteins which may similarly be screened and/or characterized. An important aspect of the present invention is that the proteins that are formed are designed to incorporate inteins or other cleavage producing portions of a protein. These cleavage sites allow the protein itself to be spliced and recombined to form proteins which may have a suitable activity.

Another aspect of the present invention is a method of producing proteins which includes isolating a first gene and a second gene; truncating the first and second genes in a controlled fashion; ligating the truncated first and second genes to form a recombinant plasmid; and expressing the protein which corresponds to the recombinant plasmid. In this embodiment it is preferred that the protein that is expressed include an intein and/or dimerization domain as a result of inclusion of the appropriate coding sequence in the recombinant plasmid or through post-translational modification of the protein after its expression by the recombinant plasmid.

As used herein, a desired characteristic or desired functionality may include any of the following traits: the absence of a characteristic, function or property ; a known and/or unknown function; an increase or a decrease in activity and novel or unexpected activities.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 schematically demonstrates the creation of an incremental truncation library.

Figure 2 schematically demonstrates the creation of a seamless ITCHY library.

Figure 3 schematically demonstrates the creation of a SCRATCHY library (created by shuffling two ITCHY libraries).

Figure 4 is a depiction of a parental incremental truncation plasmid and construction of an incremental truncation library using nucleotide analogs.

Figures 5A-5G show the construction of ITCHY libraries between two individual genes or gene fragments located on a single plasmid by simultaneous incremental truncation using nucleotide analogs by a method called THIO-ITCHY.

Figure 6 is an illustration of the CP-ITCHY principle.

Figure 7 shows the creation of CP-ITCHY libraries. Fig. 7a is a description of a vector (pDIM-N5) for creating CP-ITCHY libraries; Fig. 7b is an example of a CP insert and construction of a CP-ITCHY library.

DETAILED DESCRIPTION OF THE INVENTION

So that the invention described herein may be more fully understood, the following detailed description is set forth. The description is in no way meant to limit the breadth of the claims, but rather to specifically point out the novel aspects of the present invention.

The present invention relates generally to a combinatorial methodology for creating libraries of hybrid proteins independent of DNA homology. This invention allows for creating libraries of fusions at regions of homology and non-homology including regions such that internal "deletions" and/or "duplications" (from the aligned sequences or structures) are created. Embodiments of the present invention may be used to create such hybrid protein libraries. The present invention may also involve a series of combinatorial approaches that utilize the incremental truncation of genes, gene fragments or gene libraries.

Generally, for incremental truncation, an exonuclease (such as Exo III) can be used to create a library of all possible single base-pair deletions of a given piece of DNA. Incremental truncation libraries (ITLs) have applications in protein engineering as well as protein folding, enzyme evolution, and the chemical synthesis of proteins. In addition, the invention provides a methodology of DNA recombination which is independent of DNA sequence homology.

For the average size gene, the separate construction of all possible one-nucleotide base truncations would require the assembly of hundreds of plasmids, a labor intensive and time consuming task. The present invention, allows the construction of a library containing all possible truncations of a gene, gene fragment or DNA library in a single experiment as depicted in Fig. 1. Many of the techniques involved in the present invention make use of recombinant DNA technology, using DNA cloning vehicles and other tools of genetic engineering in the process of making the libraries of the present invention. Many of these basic techniques are described in Maniatis, Fritsch and Sambrook in *Molecular Cloning: A Laboratory Manual*; Cold Spring Harbor Laboratory: Cold Spring Harbor, NY, 1982, which is hereby incorporated by reference in its entirety.

In an embodiment of the present invention, incremental truncation by the process of slow, directional, controlled digestion of DNA is utilized for the creation of novel fusion proteins. For example, during this digestion, small aliquots are frequently removed and the digestion quenched. Thus by taking multiple samples over a given time period a library of all possible single base-pair deletions of a given piece of DNA can be created.

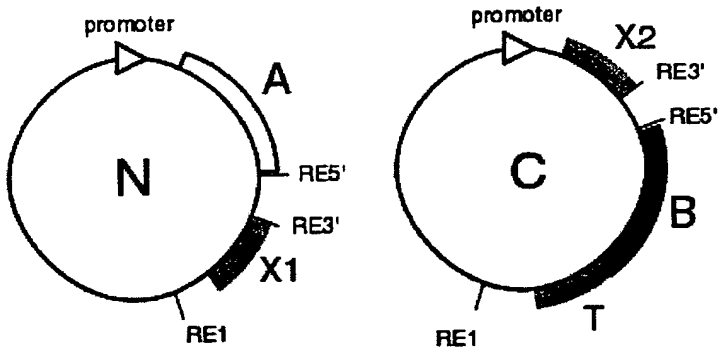
For example, Figure 1 shows the generalized procedure for incremental truncation. Incremental truncation is performed on exonuclease susceptible DNA such as linear DNA containing a gene, gene fragment or DNA library that has one end protected from digestion and the other end susceptible to digestion. This is easily accomplished, for example, by (as shown in step 1) digestion of plasmid

DNA with two restriction proteins: one that produces a 3' overhang (RE3'; which is resistant to Exo III digestion) and the other which produces a 5' overhang (RE5'; which is susceptible to Exo III digestion). Step 2 illustrates the digestion with Exonuclease III which proceeds under conditions such that the digestion rate is slow enough that the removal of aliquots at frequent intervals results in a DNA library with every one codon (or base pair) deletion. In step 3 the ends of the DNA can be blunted by treatment with a single stranded nuclease (such as S1 nuclease or Mung Bean nuclease) and Klenow so that unimolecular ligation results in the desired incremental truncation library. For some applications, additional DNA manipulations are required before recircularizing the vector.

Any enzyme that can digest DNA in a controllable, directional manner could be utilized in the methodologies described herein. In the following examples, Exo III has been used and exhibits the desired properties. Exo III has been previously shown to be useful in the creation of large truncations of linear DNA and for techniques in the sequencing of large genes. However, previous techniques utilized the digestion rate of Exo III at 37 °C (~500 bases/min), which is much too fast for purposes of incremental truncation where every one-nucleotide base deletion is desired. The fact that the digestion rate of the exonuclease can be affected by a variety of methods, such as lowering the incubation temperature, altering the digestion buffer composition, inclusion of a nuclease inhibitor or lowering the ratio of enzyme to DNA is advantageous to the present invention. Embodiments of the present invention modulate conditions affecting the digestion rate of Exo III so that the degradation is slowed, thus allowing for incremental truncation where potentially every nucleotide base can be deleted.

One theory on the evolution of enzymes posits that catalytic function arises from the interaction of protein fragments that eventually become condensed to a single gene product. The reverse of this process (also referred to as protein fragment complementation) would be to convert an existing monomeric enzyme into its functional

heterodimer. The use of incremental truncation libraries (ITL), in conjunction with a suitable screen or selection, such as utilizing an auxotrophic host or antibiotic selection, will determine all the points in the backbone polypeptide chain that can be broken. The two resulting fragments would still retain the ability to fold and associate into an active heterodimer when a functional selection mechanism is utilized. Importantly, several embodiments of the present invention will allow this process of reverse evolution to be performed in vitro in a reasonable amount of time.



Application	X1	X2	T
Protein Frag. Compl. with dimer fusion	3-frame stop dimer domain	start codon dimer domain	stop stop
Seamed ITCHY	RE2	RE2	stop
Seamless ITCHY	--------	--------	stop
Trans Inteins	intein (I_N)	intein (I_C)	stop
SCRATCHY	--------	--------	stop or in-frame reporter

Table 1

Various features of the vectors utilized for the applications of incremental truncation are shown in Table 1. In Table 1 plasmids N and C are two compatible vectors with origins of replication belonging to different compatibility groups and bearing genes coding for different antibiotic resistances. For some applications, it is advantageous that the two vectors are phagemids (*e.g.*, that they

also contain a phage origin of replication) for packaging into phage particles. The gene, gene fragment or gene library to be truncated (shown as A and B in Table 1) is positioned downstream from a promoter. The identity of some features of the vectors is shown in Table 1 and depends on the specific application of the method. The X1 and X2 segments (when used) represent the piece of DNA that the ITLs of A or B are fused to in the unimolecular ligation step. The use of 'RE' designates a unique restriction enzyme site. RE5' and RE3' indicate that digestion with the restriction enzyme produces a 5' or 3' overhang respectively. A 5' overhang is susceptible to Exo III digestion whereas a 3' overhang is not susceptible.

An illustration of an application of the principles represented in Table 1 involves dividing the gene for a protein (P) into two non-active, overlapping fragments: A (containing the N-terminus of P) and B (containing the C-terminus of P) which are cloned into vectors suitable for incremental truncation. For this experiment, X1 is a series of stop codons in all three frames, X2 is the start codon ATG, and T is a stop codon in frame with B. After linearizing the vector with restriction enzymes RE3' and RE5' and subsequent incremental truncation, unimolecular ligation results in the 3' end of the ITL of A being fused to a series of stop codons in all three frames and the 5' end of the ITL library of B being fused to a start codon. Although two-thirds of the ITL library of A will have 1-3 foreign amino acids on the end and two-thirds of the ITL library of B will be out of frame, one-third of each library will be in-frame and not code for any foreign amino acids. Crossing the ITL libraries of A and B, for example, by transforming both libraries into the same *E. coli* cells, will have each cell producing a different combination of an N-terminal fragment and a C-terminal fragment of the original protein, P. Active members of this crossed ITL library can be identified by screening or selection. This methodology has recently been successfully applied to *E. coli* glycinamide ribonucleotide transformylase.

Identifying points for functional bisection of an enzyme has applications in enzyme evolution and protein folding, since such

bisection points potentially identify ancestral fusion points as well as independent folding units. Such dissection of enzymes into smaller fragments also subverts impediments in the chemical synthesis of enzymes: enzymes too large to be chemically synthesized as a monomer can be synthesized as fragments, thus allowing the introduction of unique side chain functions. Moreover, the identification of functional structural motifs, subdomains, or domains will facilitate the construction of hybrid proteins and the creation of proteins with novel activities (*e.g.*, antibiotics with improved effectiveness). The construction of crossed ITLs of protein structural homologs illustrates one combinatorial approach to domain swapping made feasible by incremental truncation of the present invention.

Bisection of a protein in the manner described above could potentially lead to problems with association of the two fragments, particularly between structural homologs. The two protein fragments may be unable or have little tendency to associate. The addition of tight binding dimerization domains by using a dimerization motif could circumvent this issue.

This type of facilitated association of protein fragments would allow for the creation of structural-homolog heterodimers. One could imagine, for example, creating hybrid proteins such that an ITL of the catalytic machinery of one enzyme (A) is fused to one dimerization domain (X1) and a ITL of a substrate binding domain (B) is fused to a second dimerization domain (X2). Such A-X1 and B-X2 fusion libraries could then be crossed into *E. coli* cells, as described above for example, and the functional association of the two subunits A and B would be facilitated by the dimerization of X1 and X2. Although not necessary, X1 and X2 should preferably be different (*e.g.*, they form a heterodimer) so as to avoid homodimerization of A-X1:X1-A and B-X2:X2-B in lieu of heterodimerization (A-X1:X2-B). Structures such as anti-parallel helixes, parallel helix-turn-helixes and inactive intein domains may also be preferable to avoid the necessity of long linkers. This type of approach would allow scanning for novel activities across

families of proteins in one experiment, as A and B need not be a discrete genes but could be a library of family members.

One advantage to this approach would be the ability to access very large librariès (~10^{11}) if vectors N and C are phagemids and can be packaged into phage particles. Since phage infection is a very efficient method of introducing vectors into *E. coli*, the library size is limited primarily by the number of *E. coli* cells in the culture. For example, if each individual A-X1 and B-X2 library has a library size of $2x10^6$, then the crossed library of these two has a maximum library size of $4x10^{12}$. If a liter of 10^{11} *E. coli* cells is infected with phagemid containing each of the ITL-dimer libraries, and 30% of the cells become infected with both vectors, then the crossed library size is $3x10^{10}$. Although the ability to use selection on such large libraries can be problematic, such methodology still makes facile the creation of smaller, manageable libraries.

In hybrid proteins created by domain swapping, it can be difficult to predict exactly which fusion-points will produce a protein with desired properties. The use of incremental truncation in the creation of hybrid protein libraries solves this problem by a stochastic method. A novel feature of this method is that it is not dependent upon homology on the DNA level or any knowledge of the structure of either enzyme (or protein). Theoretically, all possible combinations of two genes can be created and, with the use of a suitable screen or selection, active hybrids can be identified. Variations or embodiments of this methodology, which we have termed Incremental Truncation for the Creation of Hybrid enzYmes (ITCHY), are outlined below.

Seamed ITCHY libraries are created for example, referring to Table 1 wherein X1 and X2 are identical restriction sites (RE2) and T is a stop codon in frame with B. The individual ITLs of A and B are constructed as in protein fragment complementation above (*e.g.*, linearization of the plasmid DNA with RE3' and RE5' followed by incremental truncation and recircularization). Next, the ITL of B is cloned into plasmid N bearing the ITL of A between the RE2 and RE1

sites using identical restriction sites on plasmid C. The resulting ITCHY library is seamed since it will contain the restriction enzyme site RE2 at the junction of the two gene fragments and thus code for foreign amino acids. One third of the library will have B in frame with A. If a linker is desired between the two genes, it can be included in either X1 or X2 such that it is between RE2 and the truncated gene.

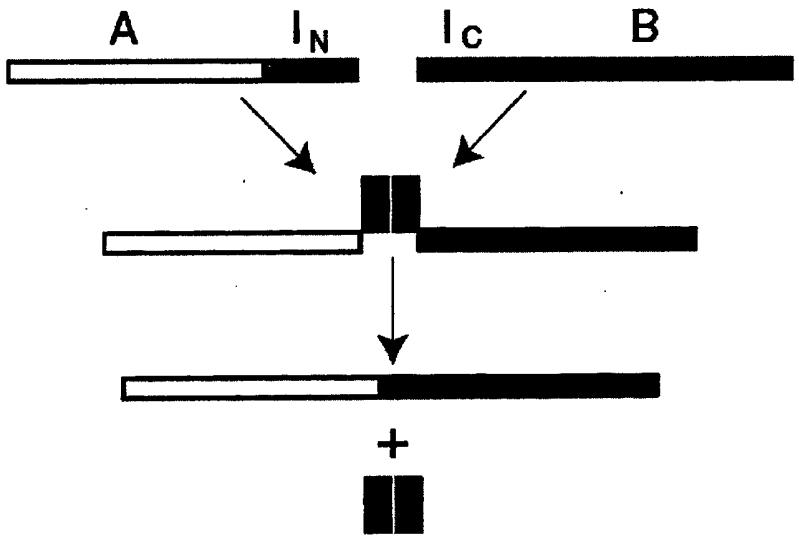
Seamless ITCHY Libraries are useful for avoiding the seam at the interface between the two genes. This method, however, depends on the cloning of fragments with one blunt end, so the library size may be less than in a seamed ITCHY.

For example, the linearized versions of vectors N and C from Table 1 are prepared by digestion with RE3' and RE5' as shown in step 4 of Figure 2. Incremental truncation proceeds as in Figure 1. In step 5 of Figure 2, the linear ITLs are digested with RE1 and the indicated fragments are isolated. In step 6 of Figure 2, ligation of the fragments containing the ITL of B into the vector containing the ITL of A proceeds by a sticky end ligation at the site of the asterix and a blunt end ligation between the truncated genes.

Generally, incremental truncation proceeds as in protein fragment complementation above, except that before the vector is recircularized, plasmids N and C are digested with RE1 (Figure 2). Vector N (containing the ITL of A) is isolated away from fragment X1 and the ITL of B is isolated from the rest of the vector C. The ITL of B is then ligated into vector N (containing the ITL of A) by a sticky/blunt ligation. The blunt end ligation is what produces the seamless fusion of the two genes. The sticky end ligation (at RE1) provides directionality and improved cloning efficiency (compared to a blunt end ligation). As in a seamed ITCHY, one-third of the library will have B in frame with A. Unlike a seamed ITCHY, a seamless ITCHY is not easily amenable to linker incorporation. For example, we have created seamless ITCHY libraries of up to 7,600,000 fusions (2,530,000 in-frame fusions) between the incremental truncation libraries of two genes. This library size is the theoretical minimum necessary to have

all possible fusions between two ITLs whose members contain between 0 and 2,757 deleted bases.

Protein splicing is a post-translational event involving precise excision of an intein fragment from precursor protein sequences. While most inteins described to date have been *cis*-inteins (encoded on one polypeptide), recently engineered and naturally occurring *trans*-inteins have been described. The ability of *trans*-inteins to fuse potentially any two polypeptides is well suited for the creation of hybrid enzyme libraries. A fusion example is shown in the diagram below.



In this diagram, for example, fusion proteins of an ITL of A and the N-intein (I_N) and of an ITL of B and the C-intein (I_C) associate in solution via the interaction of I_N and I_C. The intein heterodimer (I_N:I_C) directs the splicing reaction resulting in the joining of A to B with a native peptide bond and the release of I_N:I_C.

Generally, in this embodiment, incremental truncation is performed as in the protein fragment complementation described

above, resulting in a fusion of an ITL of A to one half of the *trans*-intein (I_N) and an ITL of B to the other half of the *trans*-intein (I_C). If desired, a linker could be incorporated so that either A or B or both are fused to a linker after incremental truncation. Both vectors (containing an ITL fused to an intein or linker-intein) could then be introduced into the same cell and hybrid proteins created *in vivo* as a result of the intein's activity. All the hybrid protein products produced using *trans*-inteins will necessarily have a one residue from the intein at the fusion point.

As in the use of dimerization domains for protein fragment complementation above, one advantage in the use of *trans*-inteins is that very large hybrid enzyme libraries are possible. These libraries would theoretically be much larger than even those made by genetic fusions above (ITCHY libraries).

The successful creation of functional hybrids between two or more genes was historically thought to require a sufficient degree of homology on the DNA level. Current methods of in vitro and in vivo recombination of genes (such as DNA shuffling) depend on the genes having a sufficient degree of homology. However, many interspecies homologs have sequence homology below that which traditional in vitro and in vivo recombination methods can be efficiently be performed. That is, on the nucleotide level, there is maybe 30-40% sequence identity. We now appreciate that proteins with little or not sequence identity, however, can have strong structural homology. It is reasonable to assume that the recombination of such genes, for example within a fold superfamily, could result in hybrid proteins with interesting and useful properties. Furthermore, it is also reasonable to assume that recombination between genes with higher homology at loci of little or no homology could result in hybrid proteins with interesting and useful properties.

For example, another embodiment of the present invention provides for a method of recombining genes that does not require any sequence identity. These recombination techniques, use as its starting point, either seamed or (preferably) seamless ITCHY

libraries as outlined above. Whereas crossover points between genes in traditional DNA shuffling are defined and confined by the regions of identity, shuffled ITCHY library crossover points are defined by the fusion-points. An ITCHY library theoretically will have all possible crossover points; thus there wold be no limitation on the location of crossover points in the resulting hybrid enzyme library. It follows then, that shuffled ITCHY libraries (which we call SCRATCHY libraries) of genes of high identity will create more diverse libraries than traditional DNA recombination methods.

With reference to Fig. 3, a SCRATCHY library can be created by making two ITCHY libraries: one library formed with gene A on the N-terminus creating A-B fusions, and one library formed with gene B on the N-terminus creating B-A fusions. Next, DNA fragments of each of the A-B and B-A fusions are isolated that are approximately the same size as the original genes. This can be done by gel electrophoresis or capillary electrophoresis after restriction enzyme digestion (and judicious location of restriction sites) or after PCR with primers near or just outside the ends of fused genes. This step is to attempt to ensure that the pool of DNA to be shuffled contains fusions at points on the primary and three-dimensional structure which are near each other (*i.e.*, limit crossover points to 'intelligent' locations). Thus, the SCRATCHY methodology is preferably done with genes A and B being roughly the same size. This DNA with 'intelligent' crossover points may then be amplified by PCR to obtain enough sample to perform DNA recombination or shuffling. The two libraries (which include A-B and B-A PCR products of approximately the same size as the original genes) are then mixed, can then be subsequently digested with DNase I, and may be followed by a method for in vitro or in vivo recombination.

Fig. 3 shows an example of non-homologous shuffling or recombination of ITCHY libraries, wherein step 7 illustrates that individual A-B and B-A ITCHY libraries are constructed, for example, as shown in Fig. 2. Step 8 illustrates that either through use of outside restriction enzymes or outside PCR primers, those members of the

ITCHY libraries which are approximately the same size as the original genes are isolated by gel or capillary electrophoresis. In step 9, these selected ITCHY library members are mixed and fragmented by digestion with DNase I as in traditional methods of DNA recombination. In step 10, reassembly of the random fragments can proceed by template switching that can result in full-length genes with multiple crossovers.

The number of hybrids appearing "in frame" will decrease exponentially with total number of crossovers. For example, the original ITCHY libraries will only have one-third of the hybrids in-frame. A resulting member of the SCRATCHY library with two crossovers will only have a 1 in 9 chance of being completely in-frame, with three crossovers having only 1 in 27 completely in-frame. This circumstance can be addressed by pre-selecting the original ITCHY libraries for hybrids in frame. For example, if gene B is fused in frame to a reporter gene with a selectable phenotype, then all in frame ITCHY library members with in-frame crossover points can be selected. The reporter gene need not be a part of the final SCRATCHY library since it can be easily removed in the PCR steps prior to DNase I digestion. We have successfully selected for in-frame fusions in two different ITCHY libraries by this method using the neomycin resistance gene as the reporter gene.

Another embodiment includes the pairing of (a) a dNTP analog that can be randomly incorporated into double-stranded DNA by a DNA polymerase and (b) an enzyme with 3' to 5' exonuclease activity that is not capable of excising the incorporated dNTP analog.

This approach provides for: (i) the creation of an incremental truncation library without requiring the labor intensive, time consuming process of taking timed aliquots during exonuclease digestion; (ii) the creation of ITCHY libraries on a single vector avoiding purification of desired fragments; (iii) the controlled incorporation of point mutations into incremental truncation or ITCHY libraries during a polymerase-catalyzed fill-in reaction; (iv) minimizing

the biases in truncation length inherent in the other embodiments previously discussed, and (v) minimizing the number of steps required and the time required to construct an incremental truncation or ITCHY library.

With reference to Fig. 4, the parental ITCHY plasmid is linearized by digestion with a pair of restriction endonucleases (RE's) that cut at unique sites in the plasmid, and thus generate a recessive 3'-termini (or flush ended termini) (Y) at the end to be truncated, and an hydrolysis-resistant termini (RE 2) including, but not limited to a recessive 5'-termini, at the other end.

Primary nuclease treatment may be carried out by an enzyme with 3' to 5' exonuclease activity, including, but not limited to Exo III, may then be used to perform the primary digestion of the linearized plasmid. The reaction conditions (such as temperature and salt concentration) are used to adjust the reaction rate. For example, at 22°C and at a salt concentration of 100 mM NaCl a digestion rate of approximately 10 nucleobases/minute results for exonuclease III.

The linearized plasmid is incubated with the 3' to 5' exonuclease to generate a single-stranded overhang. Shown as X in Figure 4C, the length of the truncated region and the digestion or cutback rate (as discussed in the previous paragraph) determine the incubation time required. In comparison to previous methods to generate incremental truncation libraries described herein, only a single timepoint must be taken in order to obtain a full range of truncated products.

The single-stranded portion of the plasmid, generated by nuclease treatment, is used as the template for the resynthesis of the complementary DNA strand. The reaction requires an enzyme with 5'-3' polymerization activity, metal ions, and nucleotide triphosphates. The nucleoside triphosphates in the reaction are preferably a mixture of the natural deoxynucleotides (dATP, dCTP, dGTP, dTTP) and nucleotide analogs (shown as S in Figure 4D) which are referred to as

spiking the reaction. The nucleotide analogs are incorporated at random during the synthesis of the complementary strand as depicted by three representative sequences shown in Fig. 4D.

The polymerization can be catalyzed by a DNA polymerase, including for example the Klenow fragment of *Escherichia coli* DNA polymerase I, or *Taq* DNA polymerase. Preferably, a polymerase that lacks 3' to 5' exonuclease activity is used. Utilizing a thermostable enzyme such as *Taq* DNA polymerase, has the advantage of reducing the formation of secondary structure within the single-stranded sequence, which could interfere with primer extension.

Preferably, metal ions, including but not limited to magnesium and manganese are presented in the primer extension reaction. A single metal ion or a mixture of two or more metal ions can be added to the reaction mixture to vary the fidelity of the primer extension according to methods known in the art.

Finally, all four natural deoxynucleoside triphosphates (dATP, dGTP, dCTP, and dTTP), as well as the nucleotide analogs (including but not limited to α-phosphothioate deoxynucleoside triphosphates) are mixed in a concentration ratio, determined by the length of the primary nuclease treatment (shown as X in Fig. 4C) so as to incorporate, on average, a single dNMP analog over the entire length of the resynthesized complementary strand. The ratio for the deoxynucleoside triphosphates to deoxynucleoside triphosphate analogs can be calculated by the following equation:

(1)

$$\frac{1}{X}\,\delta\,[C] = [S]$$

X = length of primary nuclease digestion

δ = correction factor

[C] = concentration of dNTPs

[S] = concentration of α-S-dNTPs

The correction factor δ is determined experimentally for the individual nucleotide analog that is used in the spiking reaction. The correction factor reflects the efficiency by which the nucleotide analog is utilized by the polymerase in comparison to the natural nucleoside triphosphates.

To illustrate the above equation, a primary digestion with Exo III over approximately 300 nucleotides (X = 300) would set the concentrations of the reactant as following: at a concentration of 200 μM for each dNTP ([C]) and δ= 1, the concentration of each α-S-dNTPs ([S]) would be 0.67 μM.

The reaction mixture is then incubated at a temperature appropriate for double-strand synthesis by the enzyme. The temperature therefore can, but must not necessarily, be set at the

manufacturer-recommended activity optimum, giving access to additional random mutations under suboptimal reaction conditions.

After completion of double-strand synthesis as shown in Fig. 4D, the dNMP analog-spiked linearized plasmid is incubated with an enzyme with 3' to 5' exonuclease activity (which carries out a second nuclease treatment) that is unable to hydrolyze the DNA beyond the dNMP analog, such as Exo III for example. Based on the random incorporation of the dNTP analog during the previous resynthesis of the complementary DNA strand, the hydrolysis will be terminated at the position of the nucleotide analog over the entire length of X, as shown by the three representative sequences shown in Fig. 4E, for example.

The reaction conditions for the second nuclease treatment are somewhat less critical than those of the first nuclease treatment. The RE2-site is protected from hydrolysis and the digestion by the 3' to 5' exonuclease will automatically be terminated upon encountering the dNMP analog in the DNA strand.

With reference to Fig. 4F, after the second nuclease treatment, the single-stranded portions of the plasmid are degraded upon addition of a nuclease that specifically hydrolyses single-stranded DNA, for example S1 nuclease or Mung bean nuclease, thereby generating blunt ends.

To improve the cyclization efficiency, the plasmid can be briefly incubated with a DNA polymerase, preferentially the Klenow-fragment of *E. coli* DNA polymerase I, in the presence of metal ions and the natural deoxynucleoside triphosphates.

The blunt-ended truncated library can then be recyclized as shown in Fig. 4G, using chemical or enzymatic methods, including for example DNA ligases such as T_4 DNA ligase, at the conditions recommended by the manufacturers.

The following example demonstrates the construction of fusion protein libraries between two individual genes or gene fragments, located on a single plasmid, as shown in Fig. 5A, by simultaneous incremental truncation.

Under these specific conditions, the linearization can be achieved with only a single restriction endonuclease that generates a recessive 3'-termini or a flush-ended termini as symbolized by "Y" in Fig. 5B.

Upon incubation with a 3' to 5' –exonuclease, (for example Exo. III) gene or gene fragment A and B are hydrolyzed simultaneously over the distance X, generating a stretch of single-stranded DNA. The length of X can be controlled by the reaction conditions, including but not limited to such elements as the enzyme, the composition of the reaction buffer, the reaction temperature, and the incubation period.

Resynthesis of the complementary DNA strand by a DNA polymerase, (for example the Klenow fragment of *E. coli* DNA polymerase I, or *Taq* DNA polymerase) in the presence of metal ions and a mixture of natural deoxynucleoside triphosphates and nucleotide analogs (symbolized [S) in the appropriate ratio (see Table 2 for guidelines to determine the calculation of the required nucleotide analog concentration) leads to the random incorporation of nucleotide analogs in both directions over the entire stretch (X) of the resynthesized complementary DNA, as shown in Fig. 5D. As mentioned previously in the general procedure, a series of variables can be used to further randomize the DNA at this stage, including such elements as the type of DNA polymerase, the reaction buffer composition, the metal ion(s) present in the reaction mixture, and the reaction conditions in general.

After completion of double-strand synthesis (Fig. 5D), the dNMP analog-spiked linearized plasmid is incubated with an enzyme with 3' to 5' exonuclease activity that is unable to hydrolyze the DNA beyond the dNMP analog (*e.g.*, Exo III). Based on the random

incorporation of the dNTP analog during the previous resynthesis of the complementary DNA strand, the simultaneous hydrolysis in both directions will be terminated at the position of the nucleotide analog, as depicted by three representative sequences shown in Fig. 5D.

The reaction conditions for the second nuclease treatment represented in Fig. 5E, are less critical. The digestion by the 3' to 5' exonuclease will automatically be terminated upon encountering the dNMP analog in the DNA strand.

Following the second nuclease treatment, all single-stranded portions of the plasmid are degraded upon addition of a nuclease that specifically hydrolyses single-stranded DNA, such as S1 nuclease or Mung bean nuclease (Fig. 5F).

To improve the cyclization efficiency, the plasmid can be briefly incubated with a DNA polymerase, preferentially the Klenow-fragment of *E. coli* DNA polymerase I, in the presence of metal ions and the natural deoxynucleoside triphosphates.

The blunt-ended truncated library is recyclized using chemical or enzymatic methods, including but not limited to DNA ligases, preferentially T4 DNA ligase, at the conditions recommended by the manufacturers (Figure 5G).

In an additional example, the spiking of PCR may be carried with a parental DNA target which is amplified with 5' and 3' outside primers in the presence of dNTPs and a dNTP analog, using a DNA polymerase (including but not limited to *Taq* DNA polymerase) preferentially with no exonuclease activity. The ratio between dNTP and dNTP analog is such that on average only a single dNTP analog is incorporated per region to be truncated (as presented in example 1). Reaction conditions (for example reaction buffer composition, reaction temperature, metal ions (for example magnesium and manganese)) can be varied to affect the fidelity of the primer extension and lead to

customizable levels of random mutagenesis during amplification according to methods known in the art.

A unique restriction site that will afford protection to truncation is located at the end which is not to be truncated of the PCR product. Following restriction digestion with said restriction enzyme, the amplification product is incubated with an enzyme with 3' to 5' exonuclease activity that is unable to hydrolyze the DNA beyond the dNMP analog (for example exonuclease III). Alternatively, it may be desirable for the truncation be performed simultaneously from both ends if for example the restriction enzyme digestion is omitted.

The single-stranded portion of the amplification product is degraded with nuclease that specifically hydrolyzes single-stranded DNA, for example S1 nuclease or Mung Bean nuclease generating blunt ends. To further increase the ratio of blunt ends, the amplification product can be briefly incubated with a DNA polymerase, preferentially the Klenow fragment of *E.coli* DNA polymerase I, in the presence of metal ions and the natural dNTPs.

The fragment library may then be cloned into a suitable vector, which may or may not contain a previously prepared DNA library, according to methods known to the art.

A further embodiment of the present invention provides a method called circular permutated ITCHY (CP-ITCHY). CP-ITCHY is a modification of previously described ITCHY that offers a number of advantages over ITCHY. The general principle of this method is represented in Fig. 6. The two genes (gene 1 and gene 2) are of approximately the same length (N). It is desired to make a library of all possible fusions between N-terminal fragments of gene 1 and C-terminal fragments of gene 2, at or near where the two genes align. The region chosen to make the fusions is between A and A+x. A vector is constructed in which between the indicated fragments of the two genes (1 to A+x of gene 1 and A to N of gene 2) is a piece of DNA (CP-insert) of length x with a unique restriction site y bases from the

fragment of gene 1. If the vector is opened up at this unique restriction site and the DNA is truncated with Exo III in both directions for the amount of time necessary to truncate x bases, truncation will arrive at (A+x+y)-x = A+y in gene 1 and (A-(x-y))+x = A+y in gene 2. If the DNA of length x is a library of a this restriction site located randomly between y=0 and y=x, then truncation of this library for x bases in each direction will result in a library of all possible fusions between gene 1 and 2 between A and A+x at or near where the two genes align.

Between the two overlapping fragments of the two genes to be fused is located a piece of DNA (CP-insert) of length equal to the overlap in the two fragments. The CP-insert has a unique restriction site randomly located within. This restriction site is the start of truncation in both directions.

A sample vector for creating CP-ITCHY libraries is shown in Fig. 7a. The vector has an antibiotic resistance gene (ampicillin; Ap) as well as the two gene fragments (in this example, PurN[1-202] and GART[20-203]) cloned downstream of a suitable promoter (lac P/O). Between the two gene fragments is located a unique restriction enzyme site that produces blunt ends (*EcoR*V). This will be the site of the insertion of a circularly permuted piece of DNA.

The methodology for creating the CP-insert and the CP-ITCHY library are described in Fig. 7b. The CP-ITCHY library is created by amplifying by PCR a piece of DNA equal in length to the overlap between the two gene fragments and creating a unique restriction site at both ends (in this case *Xba*I) and cloning this DNA fragment into a vector such as pUC19. The DNA is excised from pUC19 using XbaI and treated with ligase under dilute conditions such that a significant amount of closed circular DNA is formed. The closed circular DNA is linearized at random sites by digestion with very dilute amounts of Dnase I. The randomly linearized DNA is repaired using a DNA polymerase and DNA ligase and cloned into the *EcoR*V site of pDIM-N5 by blunt end ligation. This library of *Xba*I sites between the two gene fragments is the source DNA for incremental

truncation. The library is digested with *Xba*I to linearize the vector and digested with Exo III for the length of time described in Fig. 6. The single stranded overhangs are removed by Mung Bean nuclease, the ends are blunted with Klenow and ligation under diluted conditions results in the CP-ITCHY library.

The principle advantages of CP-ITCHY over ITCHY are (a) one vector instead of two, (b) truncation in both directions simultaneously, (c) does not require extensive time point sampling, (d) biases the library considerably towards fusions at or near where the sequences align (*i.e.*, where it is most likely to produce active fusions), and (e) considerably shortens the protocol by removing steps such as extracting DNA from agarose electrophoretic gels.

Various kits that are generally useful for producing or constructing the various libraries and/or hybrid proteins described herein are also provided. Additional kits may be useful for making incremental truncation libraries and/or for producing a library of truncated recombinant plasmids. Useful components of such kits may include any or all of the following components:

(a) a purified exonuclease reagent (such as Exo III) capable of unidirectionally digesting nucleotide bases in a target DNA fragment;

(b) a recombinant vector molecule such as a plasmid or a bacteriophage vector, particularly useful ones may include pDIMN2, pDIMC8, pDIMN5, pDIMN6, pDIMC9 some of which are described in Ostermeier M, Shim JH, and Benkovic SJ, *Nat. Biotechnol.*, 1999 Dec;17(12):1205-9; which is incorporated herein by reference in its entirety;

additional useful features of the recombinant vector molecules as described in (b) above, may include restriction sites, potential sequencing primer binding

sites, a multiple cloning site, an antibiotic resistance marker, and/or a regulatable promoter;

(c) a single-strand specific nuclease enzyme such as Mung bean nuclease or S1 nuclease;

(d) buffer solutions useful in the kits may include exonuclease digestion buffers, a single-strand specific nuclease digestion buffer, a single-strand specific nuclease termination buffer, a DNA polymerase buffer and/or a DNA ligase buffer;

(e) polymerases useful in the kits may include a DNA polymerase such as Klenow fragment or Taq DNA polymerase;

(f) further elements of the kits may include a mixture of dNTPs at a specific concentration; nucleotide analogs; a DNA ligase such as T4 DNA ligase or other suitable ligase; and a suitable host for selecting a protein of desired functionality.

While the foregoing has been set forth in considerable detail, the examples are presented for elucidation and not for limitation. Modifications and improvements, including equivalents, of the technology disclosed above which are within the purview and abilities of those in the art are included within the scope of the claims appended hereto. It will be readily apparent to those skilled in the art that numerous modifications, alterations and changes can be made with respect to the specifics of the above description without departing from the inventive concept described herein. Accordingly, all such variances should be viewed as being within the scope of the present invention as set forth in the claims below.

What is claimed:

1. A method of protein identification comprising:

isolating a parent gene of interest;

truncating said parent gene in a progressive and controlled manner to form a truncated gene;

expressing a peptide corresponding to said truncated gene; and

selecting the protein based upon a predetermined characteristic.

2. The method of claim 1, wherein said pre-determined characteristic is selected from the group consisting of antibiotic resistance and growth on an auxotrophic host.

3. The method of claim 1, wherein said selected protein is screened for activity.

4. The method of claim 1, wherein said truncating involves exonuclease activity.

5. The method of claim 3, wherein said activity is the capacity of the protein to stimulate, inhibit, or modify at least one biologically active compound, said biologically active compound being selected from the group consisting of hormones, neurotransmitters, adhesion factors, growth factors, and specific regulators of DNA replication and/or transcription and/or translation of RNA.

6. The method of claim 3, wherein said activity is a lack of functionality.

7. The method of claim 1, wherein said progressive and controlled manner yields a library of incrementally truncated genes.

8. The method of claim 7, wherein one member of said library differs from a second member of said library by a length of about 1 to about 50 nucleotides.

9. The method of claim 1, wherein said progressive and controlled manner is the removal of nucleotides at a rate of about 10 base pairs per minute.

10. The method of claim 1, further including producing a library of proteins formed from a plurality of differentially modified parent genes, at least one member of the library of proteins possessing the predetermined characteristic.

11. The method of claim 1, further including the incorporating nucleotides selected from the group consisting of dNTPs, dNMPs, dNTP analogs and dNMP analogs.

12. The method of claim 1, wherein said parent gene is selected from the group consisting of a gene of interest, a portion of a gene of interest, a gene fragment, a PCR product, and a mutant of said gene of interest.

13. The method of claim 8, further comprising the step of assaying the library of proteins for a desired activity.

14. A method of forming a gene for protein expression comprising:

selecting a first parent gene and a second parent gene said first gene being selected independent of homology with said second gene;

truncating said first and second parent genes to form first and second truncated genes; and

joining said first and second truncated genes to form a combined truncated gene.

15. The method of claim 14, wherein said first and second genes are substantially homologous.

16. The method of claim 14, wherein said first and second gene are truncated in a controlled manner.

17. The method of claim 14, wherein said controlled manner is time and temperature dependent.

18. The method of claim 14, wherein said controlled manner is removal of a nucleotide at a rate of less than 10 base pairs per minute.

19. The method of claim 14, wherein said first and second truncated genes are substantially equivalent in size.

20. The method of claim 14, wherein said first and second truncated genes are fused to form a recombinant plasmid.

21. The method of claim 14, further including the incorporating nucleotides selected from the group consisting of dNTPs and dNTP analogs.

22. The method of claim 14, wherein the first parent gene includes a functional genetic element to facilitate fusion of said first and second parent gene.

23. The method of claim 14, further including repeating the steps of isolating, truncating and joining to form a library of

combined truncated genes, said library including a plurality of different combined truncated genes.

24. The method of claim 23, wherein said library of combined truncated genes is further truncated and shuffled to form a library of randomly combined truncated genes.

25. The method of claim 24, wherein said library of combined truncated genes are used to produce a library of proteins, said library of proteins including cleavage sites for recombination to form hybrid proteins.

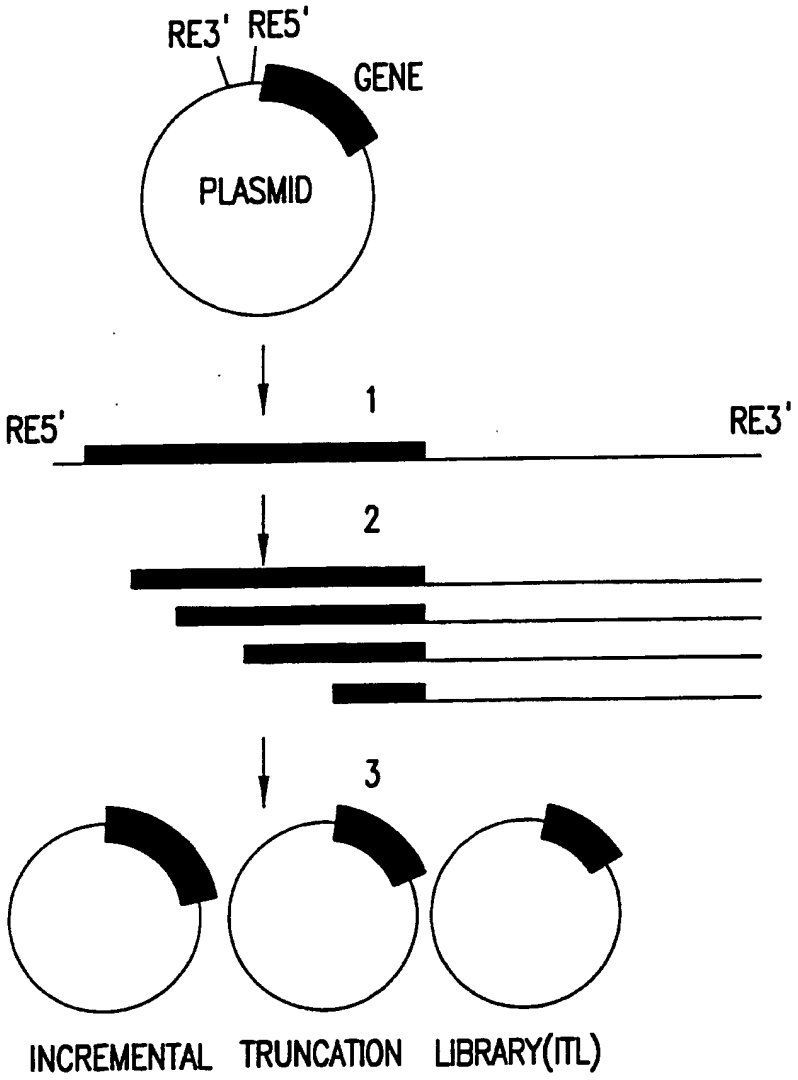

RE3'
RE5'
GENE
PLASMID
RE5'
1
RE3'
2
3
INCREMENTAL
TRUNCATION
LIBRARY(ITL)

FIG.1

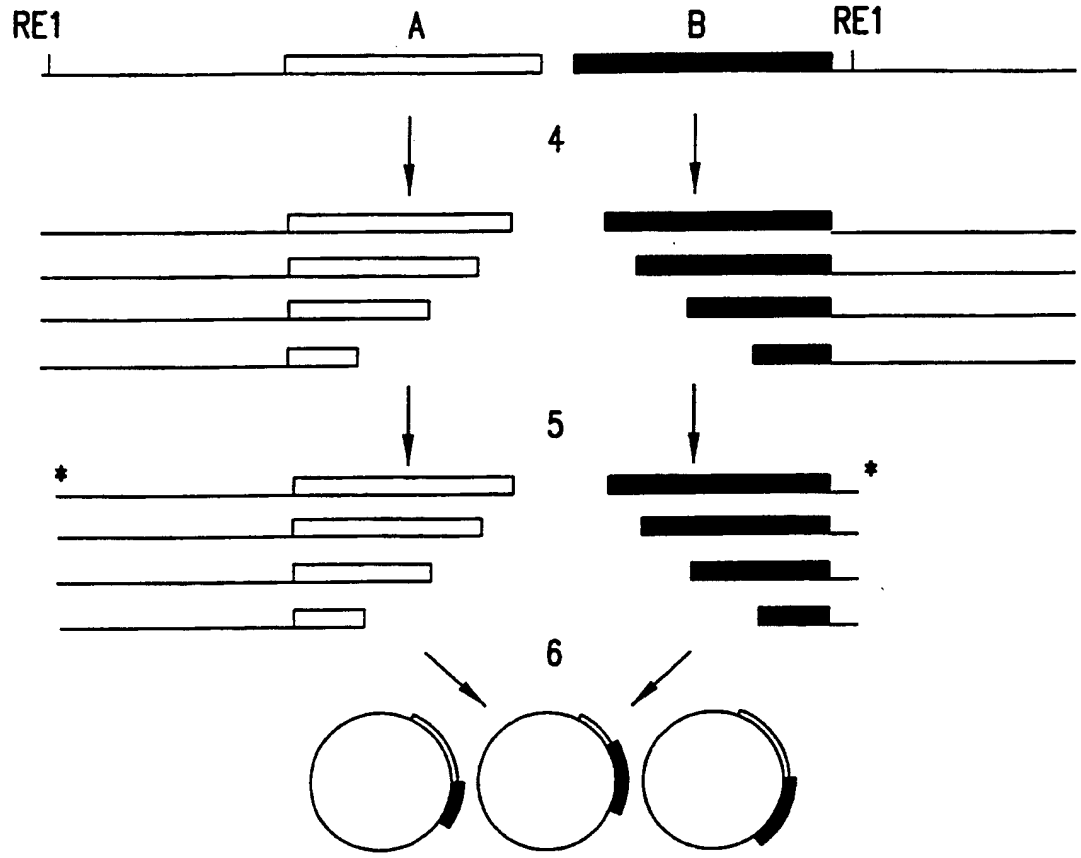


FIG.2

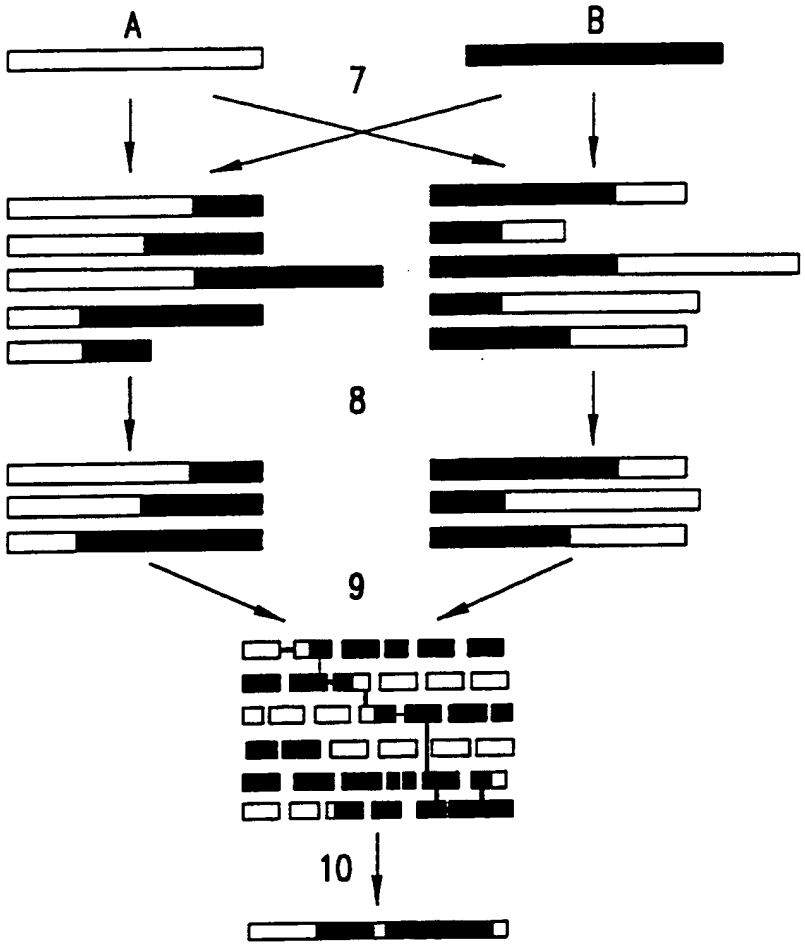


FIG.3

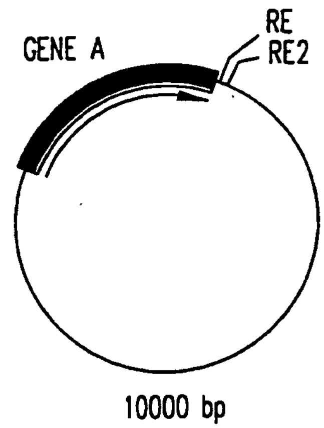


FIG.4A

LINEARIZATION

FIG.4B

PRIMARY NUCLEASE TREATMENT

FIG.4C

SPIKING

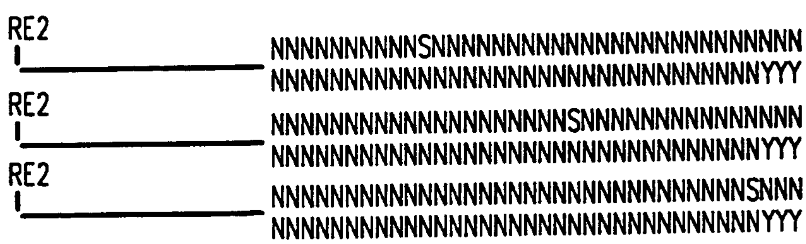


FIG.4D

SECOND NUCLEASE TREATMENT

FIG.4E

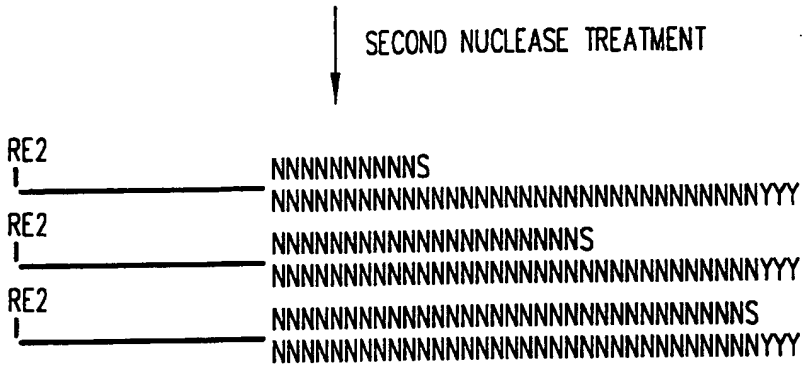


FIG.4F

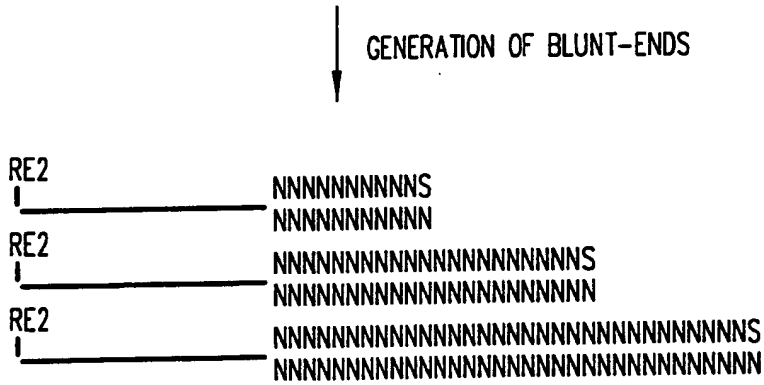


FIG.4G

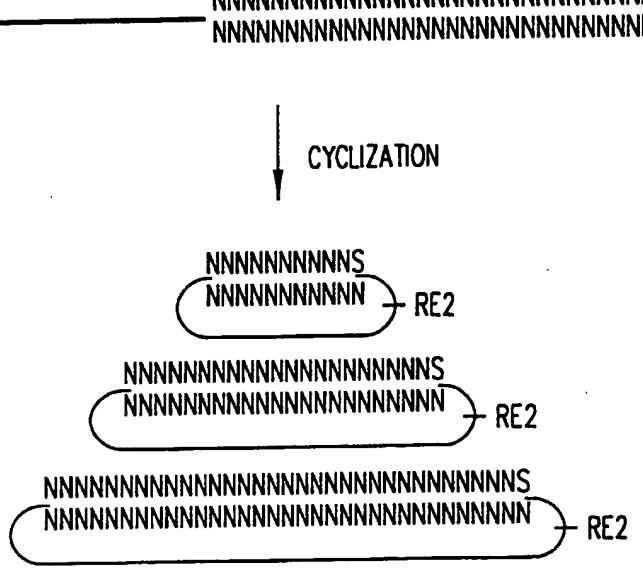


FIG.5A

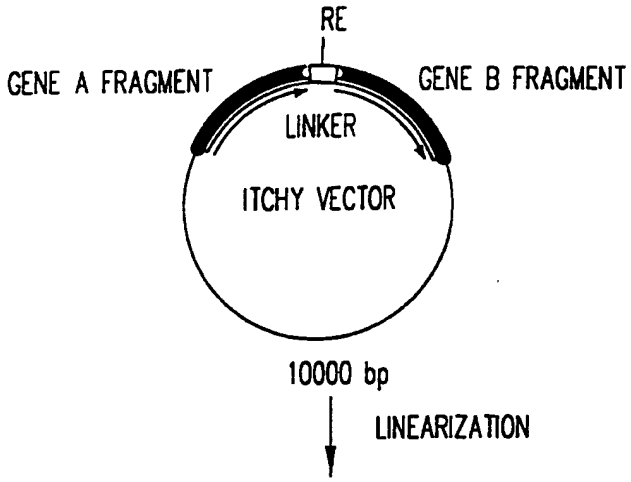


FIG.5B

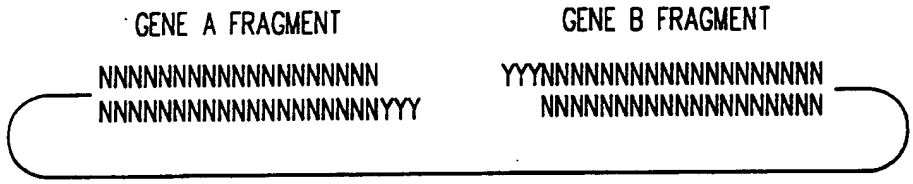


FIG.5C

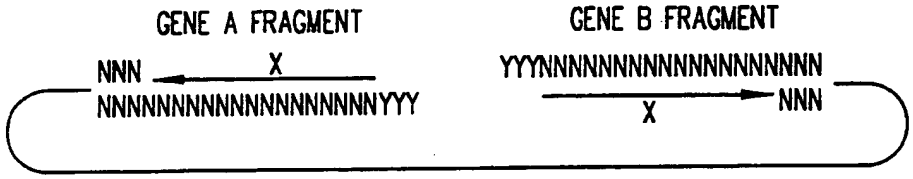
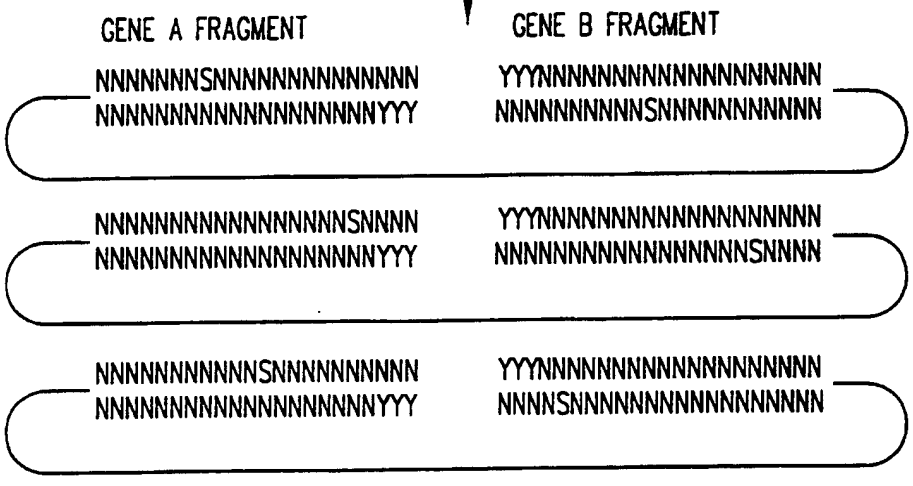


FIG.5D



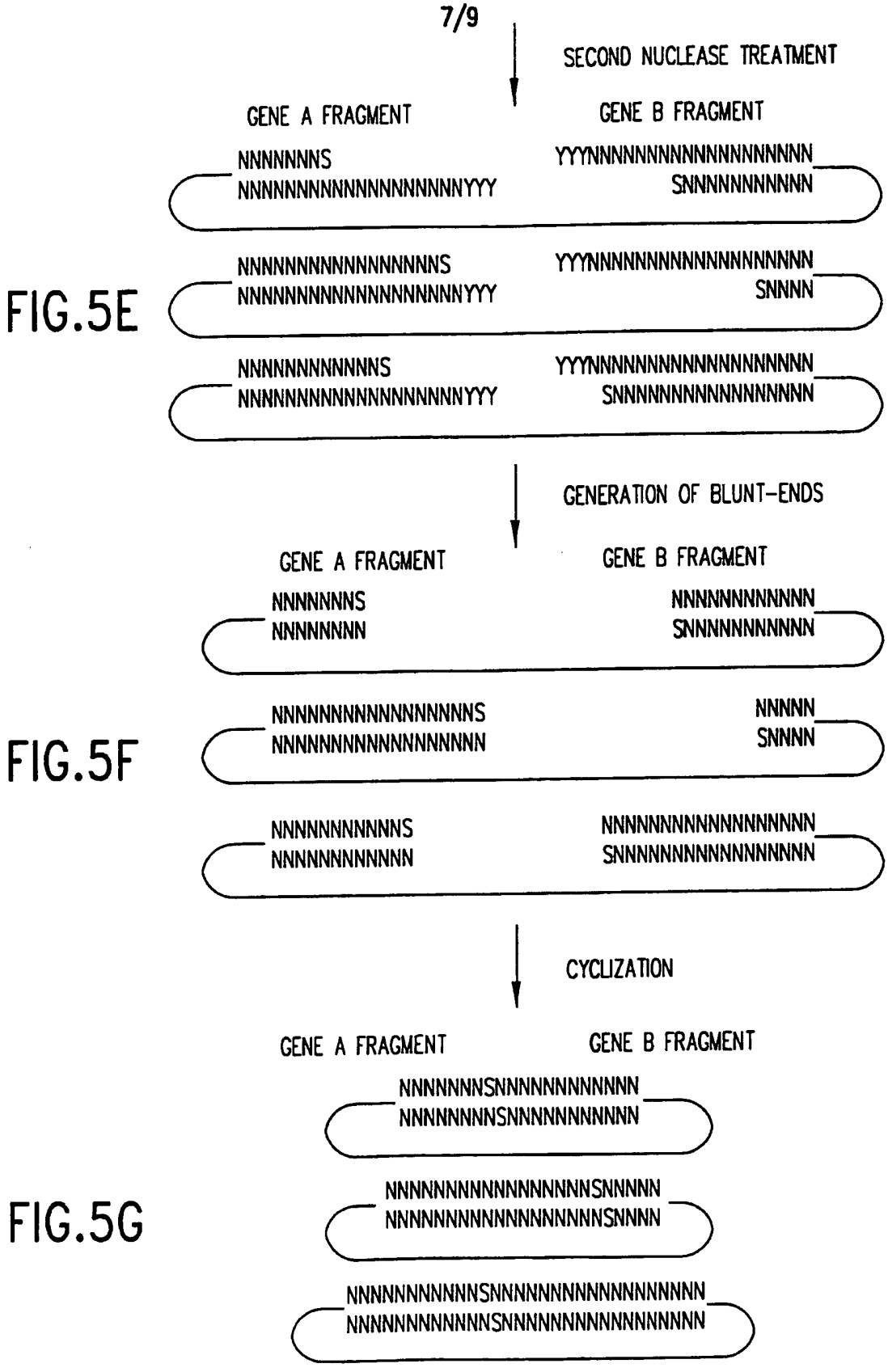


FIG.5E

FIG.5F

FIG.5G

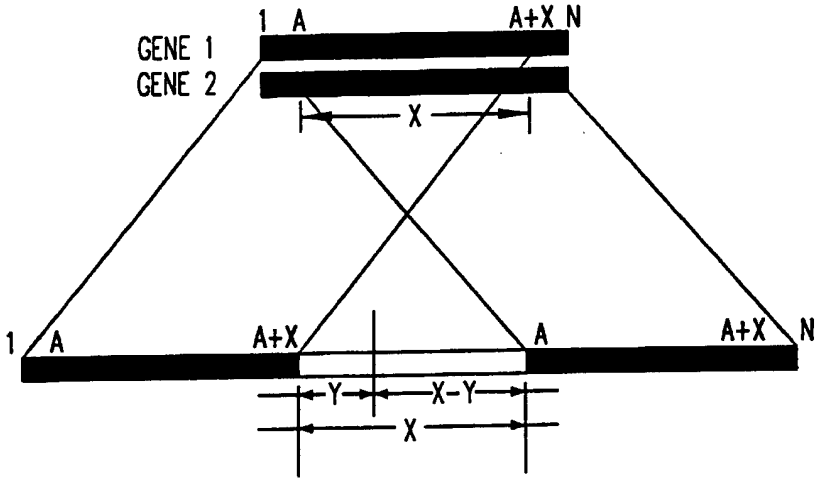


FIG.6

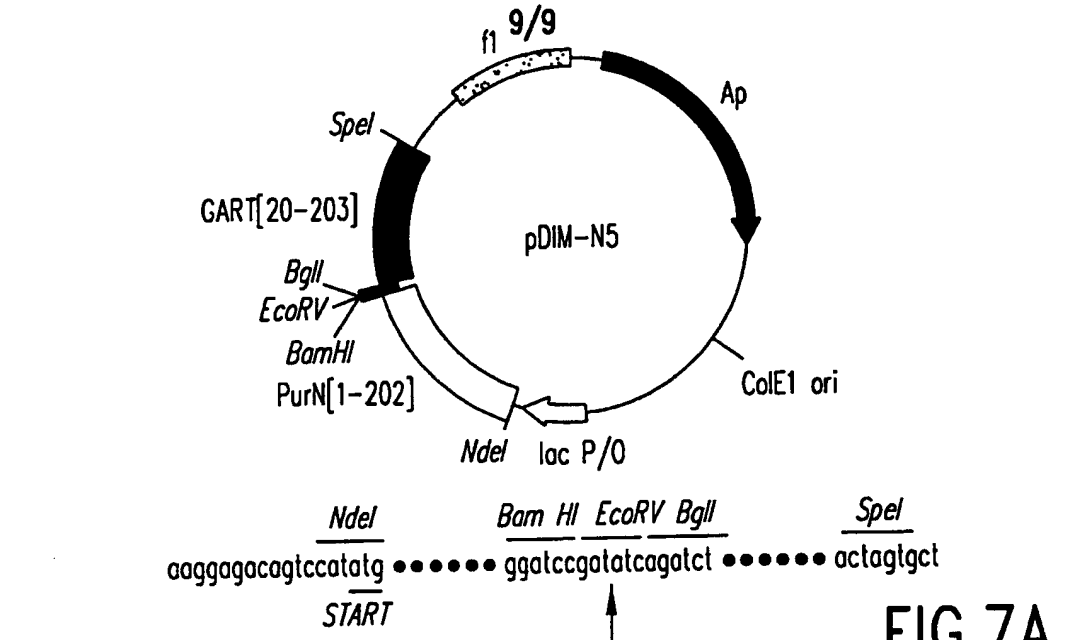

f1
Ap
SpeI
GART[20-203]
pDIM-N5
BglI
EcoRV
BamHI
PurN[1-202]
ColE1 ori
NdeI
lac P/O
NdeI
Bam HI EcoRV BglI
SpeI
aaggagacagtccatatg ●●●●●● ggatccgatatcagatct ●●●●●● actagtgct
START

FIG.7A

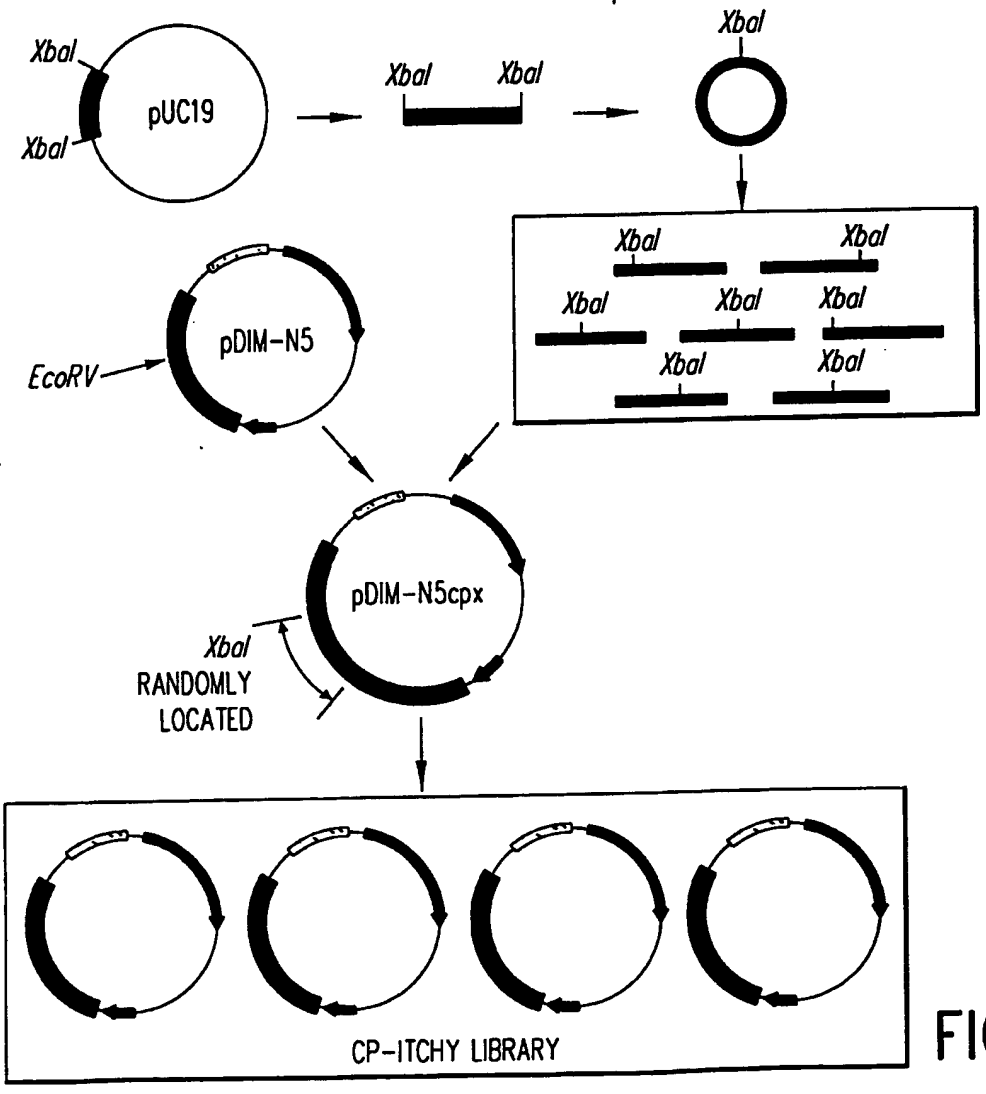

XbaI
pUC19
XbaI
XbaI
XbaI
XbaI
XbaI
XbaI
XbaI
XbaI
XbaI
pDIM-N5
EcoRV
pDIM-N5cpx
XbaI
RANDOMLY
LOCATED
CP-ITCHY LIBRARY

FIG.7B

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US00/13813

A. CLASSIFICATION OF SUBJECT MATTER

IPC(7) :G01N 33/53, 33/543; C12N 15/00; C07H 21/02, 21/04

US CL :Please See Extra Sheet.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/320.1, 7.6; 436/518; 536/23.1, 23.2, 23.4; 935/9, 10, 22, 23, 52, 79, 80

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

NONE

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

WEST, EAST, MEDLINE, SCISEARCH, BIOSIS

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	OSTERMEIER et al., "Combinatorial Protein Engineering by Incremental Truncation," Proc. Natl. Acad. Sci. USA. March 1999. Vol. 96, pages 3562-3567.	1-25
Y	HULTMAN et al. "Solid-Phase Cloning to Create Sublibraries Suitable for DNA Sequencing," Journal of Bacteriology. June 1994. vol. 35, pages 229-238.	1-25

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search

11 AUGUST 2000

Date of mailing of the international search report

07 SEP 2000

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231

Facsimile No. (703) 305-3230

Authorized officer

P. PONNALURI

Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet) (July 1998)*

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US00/13813

A. CLASSIFICATION OF SUBJECT MATTER:
US CL :

435/320.1, 7.6; 436/518; 536/23.1, 23.2, 23.4; 935/9, 10, 22, 23, 52, 79, 80